



Universiteit
Leiden
Governance and Global Affairs



Comparative Analysis of GPTs and Open-Source LLMs for Annotating Parliamentary Discourse

Bastián González-Bustamante

Leiden University

✉ b.a.gonzalez.bustamante@fgga.leidenuniv.nl

Presentation at the XXVIII World Congress of Political Science
Seoul, South Korea, July 12–16, 2025

Motivation and Research Questions

Motivation and Disclaimer

 **Why incivility and policy agenda matters?** Opportunities for MPs to question government ministers. Democratic accountability and dynamics between government and opposition.

 **Gladiatorial contest.** The questioning could have a “combative” nature and garner media attention.

 **Proof-of-concept.** ParlaMint data is *forthcoming*. This presentation's evidence is drawn from (own) prior studies on policy agenda and incivility benchmarks.



Artwork by DALL-E 3 model

Research Questions



Artwork by DALL-E 3 model

RQ1. How well do general-purpose LLMs' zero-shot capabilities transfer to political domains?

RQ2. What trade-off arises between proprietary GPTs (also reasoning models) and open-source models?

RQ3. How does task complexity (policy topics vs toxicity) influence those trade-offs?

Data and Tasks Already Benchmarked

Ground-Truth Policy Agenda

Denmark

15 101 bills
1953–2016

NLD

4 684 bills
1981–2009

UK

6 169 Acts
1911–2015

France

3 069 laws
1979–2013

Hungary

8 220 bills
1990–2022

Italy

4 554 laws
1983–2013

Brazil

2 449 laws
2003–2014

Spain

2 256 laws-decrees
1980–2018

Note. We split the samples in a proportion of 70/15/15 (stratified by major topic) for training, validation, and testing for future fine-tuning jobs. The samples correspond to ground-truth data of the Comparative Agendas Project.

Ground-Truth Toxicity

Arabic

5000 tweets
manually annotated

Chinese

5000 messages
for toxic detection

English

5000 Wikipedia
comments

German

5000 Twitter and
Facebook comments

Hindi

5000 Twitter and
Facebook comments

Russian

5000 comments on
social network OK

Spanish

5000 messages
for toxic detection

Note. We split the samples in a proportion of 70/15/15 for training, validation, and testing in case of fine-tuning jobs. The samples correspond to ground-truth data prepared for [CLEF TextDetox \(2024\)](#).

79/109 LLMs

Multiclass/Binary

Run **620/807 times** under different conditions (e.g., parameters, API/local, temperature, datasets/language)



SOTA closed-source LLMs

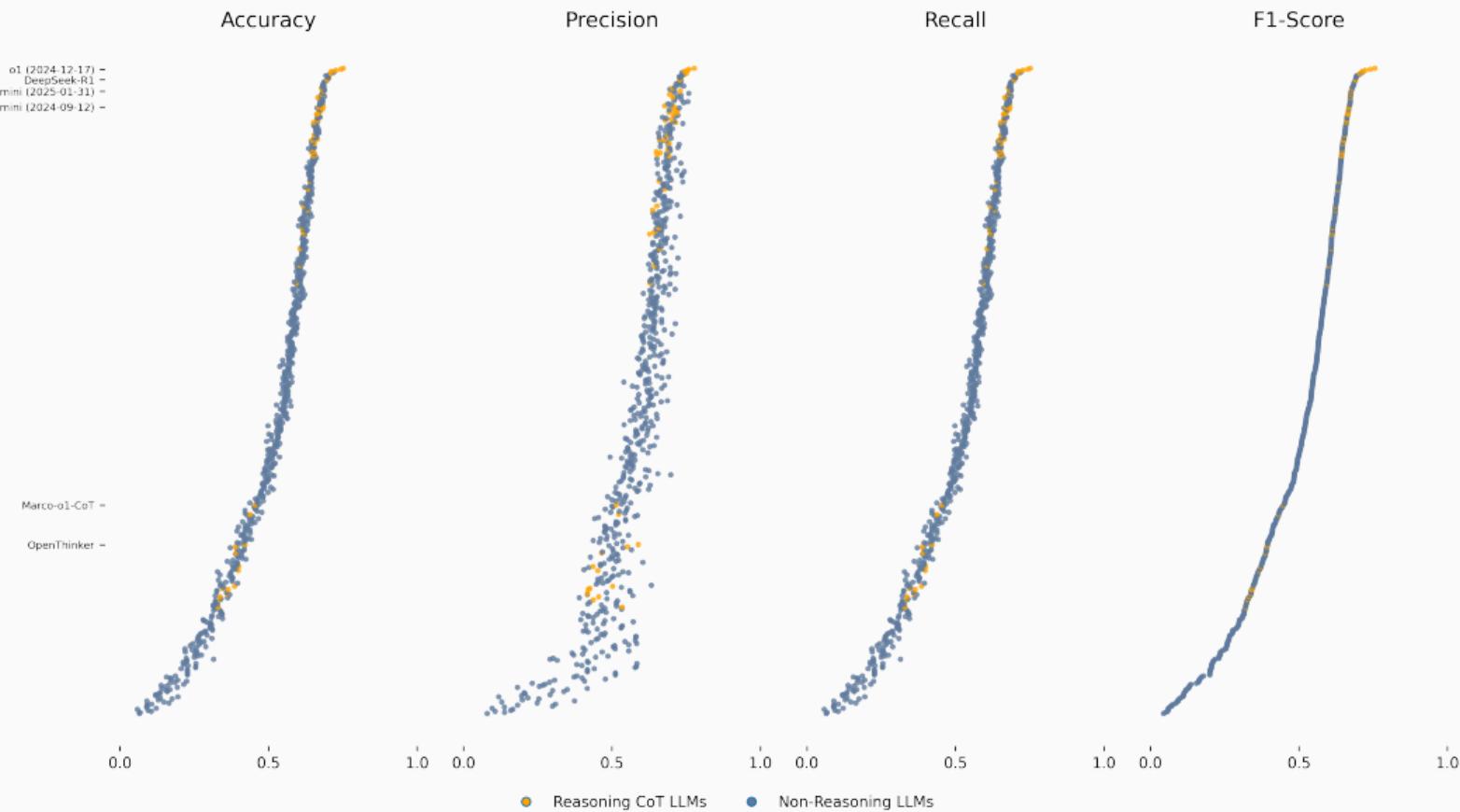
o4-mini, o3-mini, o1, GPT-4.1, GPT-4.5-preview, Grok 3 Beta, Claude 4 Opus, among others



SOTA open-source LLMs

Llama 4 Maverick (400B) and Scout (107B), Mistral 3.1 (24B), Llama 3.3 (70B), DeepSeek-R1 (671B), DeepSeek-V3 (671B), among others

Benchmarking Policy Agenda



Top Performing Models Policy Agenda

0.76

**Maximum
F1-Score**

Achieved by OpenAI
reasoning models

0.69

**Top Open-
Source Models**

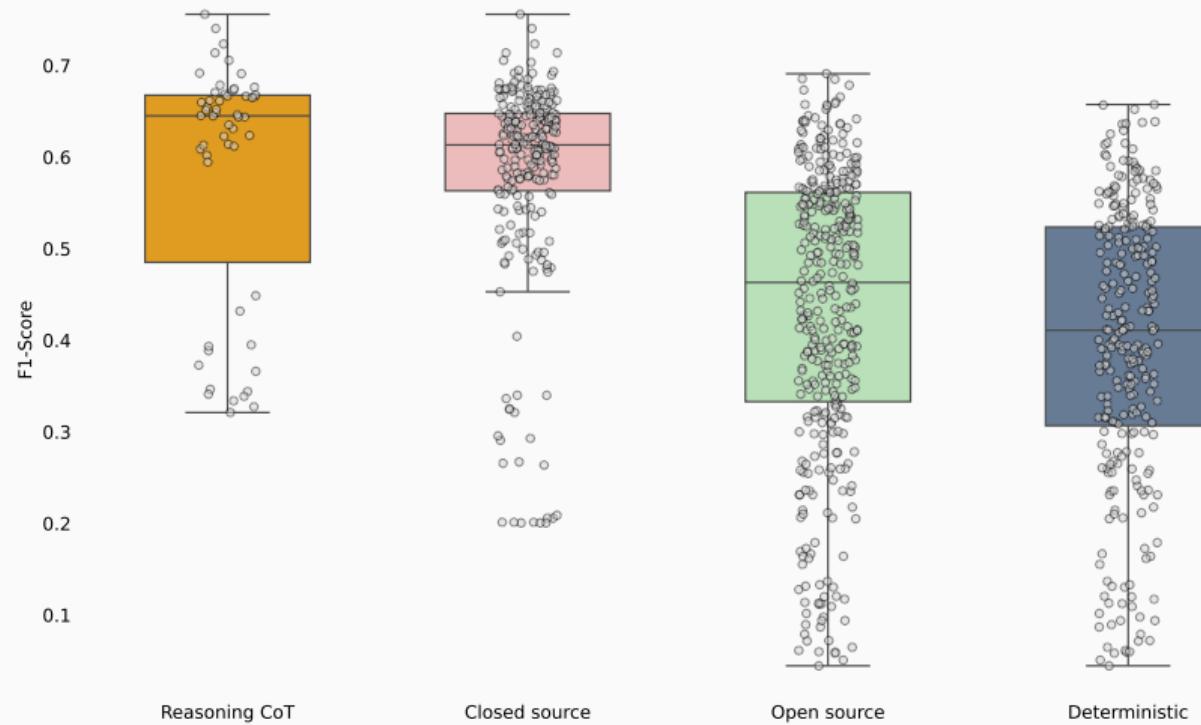
DeepSeek-R1 and V3 (671B)
and Llama 3.1 (405B)

0.54

**Mid-Range
Performance**

Gemini 1.5 series,
GPT4o-mini, Mis-
tral models, Qwen 2.5

F1-Score Across Reasoning, Source and Deterministic Setup



Comparing with Fine-Tuned Models

Language	Best LLM	F1-Score	Fine-Tuned	F1-Score	Δ Val	Δ Best LLM
Danish	GPT-4.5	0.679	Babel Machine	0.925	+0.065	+0.246
Dutch	o1	0.724	Babel Machine	0.906	+0.066	+0.182
English	o1	0.706	Babel Machine ft-XLM-RoBERTa	0.869 0.810	-0.031 -0.009	+0.163 +0.104
French	o1	0.714	Babel Machine	0.821	-0.029	+0.107
Hungarian	GPT.4-5	0.672	Babel Machine	0.751	-0.099	+0.079
Italian	o1	0.675	Babel Machine	0.930	+0.120	+0.255
Portuguese	o1	0.651	Babel Machine	0.867	-0.063	+0.216
Spanish	o4-mini	0.756	Babel Machine	0.916	+0.066	+0.160

Note. All estimates are weighted F1-scores obtained on our fixed held-out test set. The columns Δ Val and Δ Best LLM indicate: (i) the change relative to the best result on the model's own validation set; and (ii) the change relative to the strongest zero-shot LLM, respectively.

How much data leakage is present here? Probably considerable, so that the results may be inflated. However, the potential advantages of fine-tuning still seem to outweigh in-context learning, even for BERT-like models.

Key Findings Meta-Analysis Policy Agenda

+8.9%

**Reasoning
Advantage**

Chain-of-thought capabilities boost F1-score by almost 9 points

-7.8%

**Reproducibility
Penalty**

Deterministic deployment reduces F1-score by about 8 points, but ensures consistency

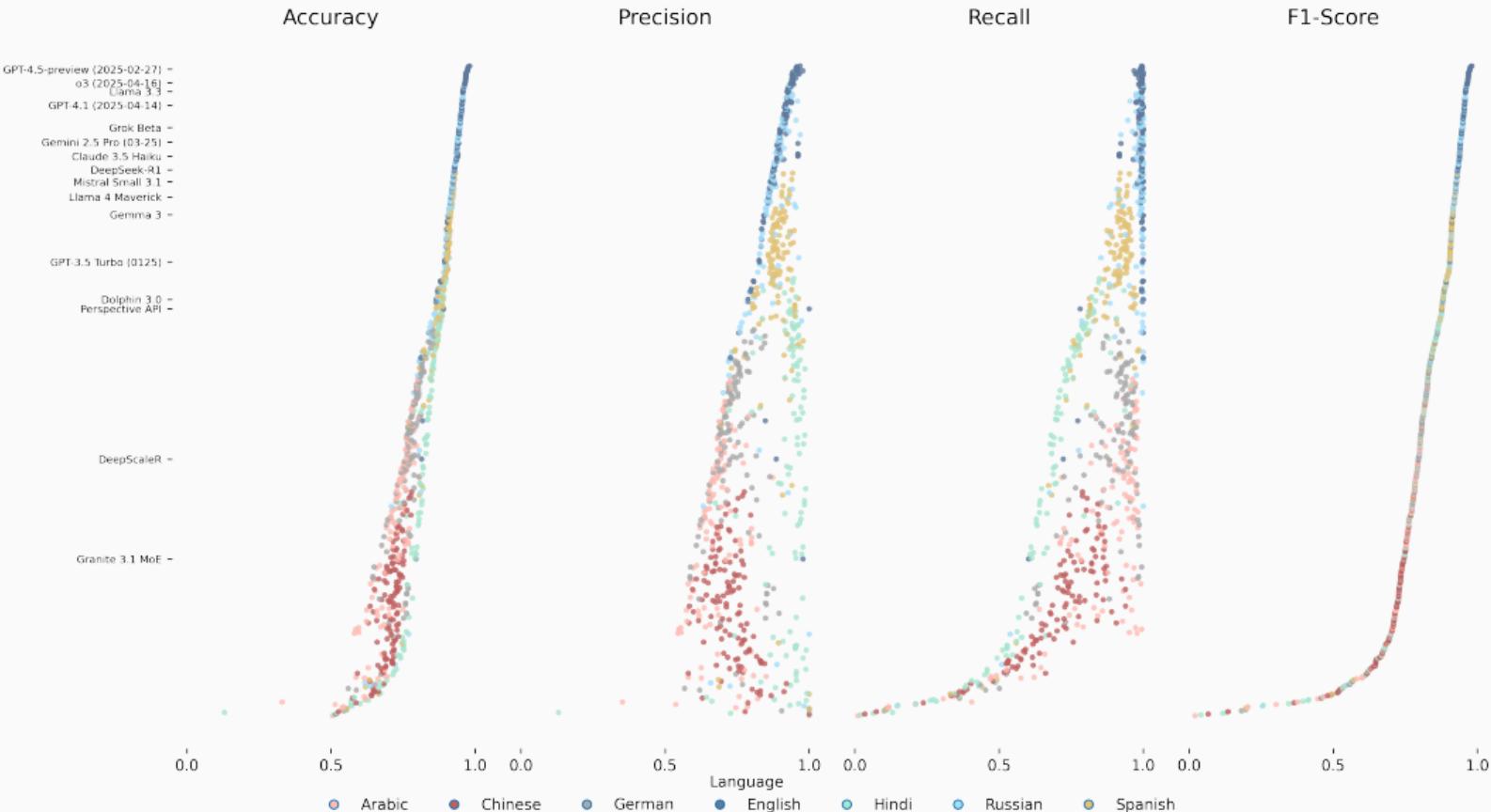
$\geq 10\%$

**Fine-Tuning
Advantage**

Supervised transformers outperform zero-shot LLMs by 10+ points

We did not find  openness penalty

Benchmarking Toxicity



Top Performing Models

Arabic

0.83 F1-Score

o3 · o1 · GPT-4.5

Chinese

0.78 F1-Score

GPT-4.1 · GPT-4o · Grok 3

English

0.98 F1-Score

Granite 3.2 ·
OLMo 2 · GPT-4.5

German

0.86 F1-Score

GPT-4.1 · o3 · GPT-4.5

Hindi

0.89 F1-Score

Mistral Saba · Mistral
Medium 3 · Gemma 2

Russian

0.96 F1-Score

o3 · GPT-4.1 nano
· Claude 4 Opus

Spanish

0.93 F1-Score

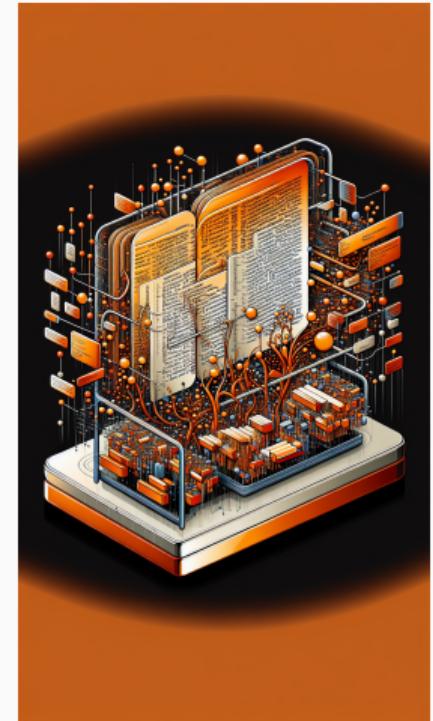
GPT-4.5 · o3 · GPT-4.1

Comparison to Perspective API

G Perspective API performance. Competitive accuracy but abysmally low recall scores, resulting in low F1-scores across languages.

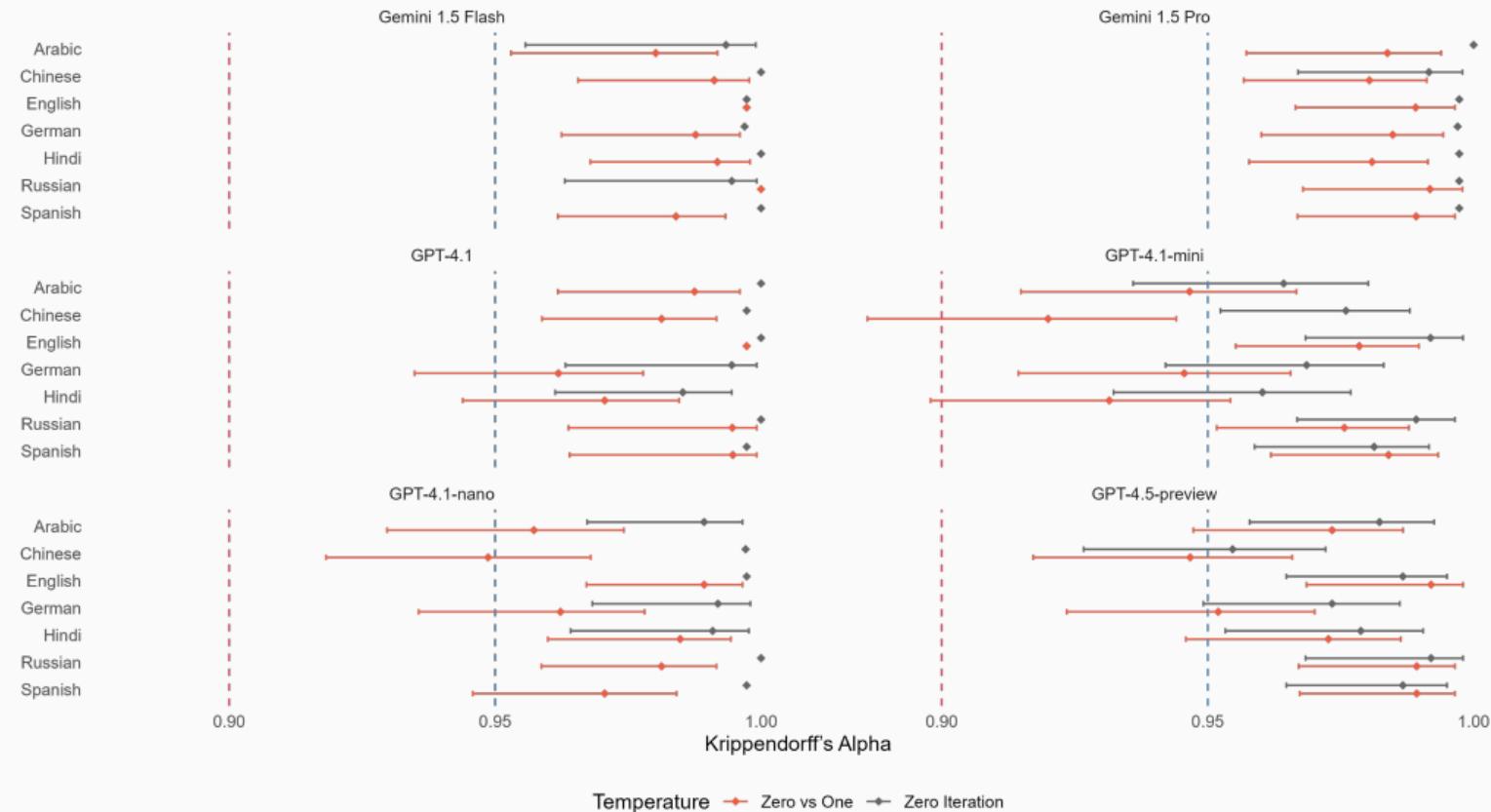
📊 Performance gap. Even weakest LLMs outperform Perspective by about 0.20 points in F1-score for Arabic and Chinese.

💻 Technological shift. Highlights transition from classical transformer pipelines to modern LLMs for toxicity classification.



Artwork by DALL-E 3 model

Temperature Experiments



Key Findings Meta-Analysis Toxicity

+7.7%

**Resource
Advantage**

High-resource languages enjoy a 7.7 points F1 advantage

-9.8%

**Size
Penalty**

Small models ($\leq 7B$) fall 9.8 points short of larger LLMs baseline

99%

**Annotation
Reliability**

Modern private LLMs exhibit outstanding reliability even w/temperature variations

We did not find  openness or  overthinking penalties

Methodological Takeaways

Evidence-Based Recommendations

- **Reasoning advantage.** Chain-of-thought capabilities enhance performance in policy agenda classification, but do not yield gains in toxicity classification.
- **Openness penalty inconclusive.** Slight lower performance than proprietary models, descriptively. There is no significant effect in meta-analyses. However, deterministic deployment reduces F1-scores in policy agenda classification.
- **BERT family pays off in complex tasks.** Supervised transformers outperform zero-shot LLMs by +10 points in policy classification. Even the weakest LLMs in toxicity outperformed off-the-shelf options.
- **Performance vs reproducibility.** Switching from deterministic to stochastic decoding rarely alters more than 1% of labels on average in toxicity.

Bridging to ParlaMint

Parliamentary Debates Data

We shall use [ParlaMint](#), which offers data on parliamentary debates of 26 European countries and three Spanish autonomous regions from about 2015 to mid-2022 (some countries even from the 1990s). The dataset offers not only the original speeches but also corpora linguistically annotated in the original languages and machine-translated versions of the texts.

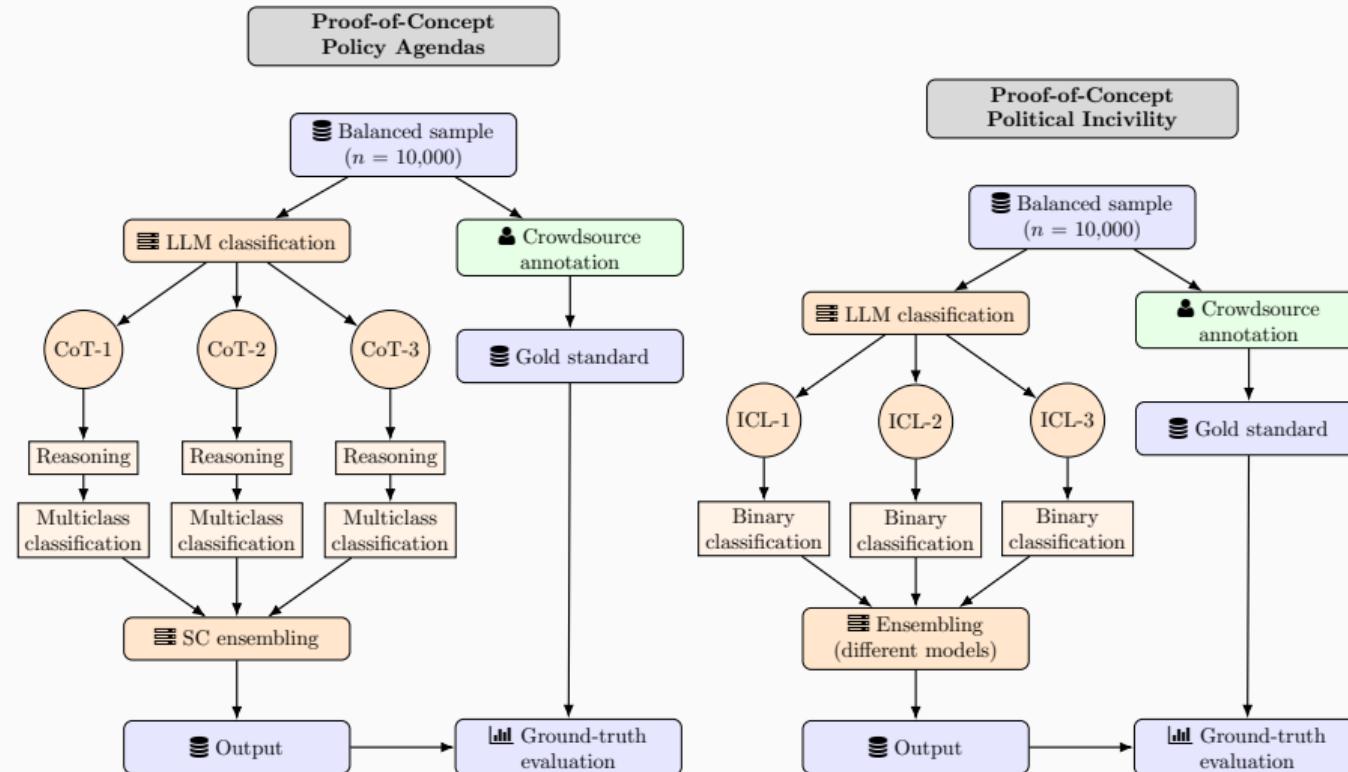


ParlaMint Corpora ~1.1B words



This implies that
~7.5M utterances
were delivered in Parliament in the case of unicameral systems or some of the relevant chambers in bicameral systems.

Proof-of-Concept Pipeline



* We may replace reasoning LLMs with fine-tuned models.

Thank you very much!

Do you have any questions?

b.a.gonzalez.bustamante@fgga.leidenuniv.nl

bgonzalezbustamante.com